Universidad
Popular del Cesar

Minería de datos y Patrones

Version 2024-1

Accuracy Estimation (Problems)

Capítulo [5]

Dr. José Ramón Iglesias
DSP-ASIC BUILDER GROUP
Director Semillero TRIAC
Ingenieria Electronica
Universidad Popular del Cesar

On Accuracy Estimation in Face Biometric Problems

Domingo Mery, Yuning Zhao and Kevin Bowyer

UNIVERSITY OF NOTRE DAME
College of Engineering
UNIVERSITY OF NOTRE DAME

FACULTAD DE INGENIERÍA
PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

Motivation

If we read in a paper on face expression recognition…

SAMPLE TITLE OF SAMPLE PAPER WITH MORE THAN ONE AUTHOR

ABSTRACT

INTRODUCTION

SAMPLE HEADING

The accuracy obtained by the proposed method was 97.3%

We could believe that 97.3% of the expressions that a person makes will be correctly recognized.

If we read in a paper on face expression recognition…

The accuracy obtained by the proposed method was 97.3%

We could believe that 97.3% of the expressions that a person makes will be correctly recognized.

However,

- how confident is this value for the dataset used in the paper?
- how generalizable is the proposed method for a wider variety of conditions?
- can this 97.3% be compared with the 98.5% reported in another paper for expression recognition experiments on the same dataset?

WHAT?

Typically we focus on the “what” elements of the dataset.

WHAT?

- What is the number of images in the dataset?
  >> Larger is better.

- What kinds of expressions were taken into account?
  >> A greater variety is generally better.

- What are the illumination conditions in the images?
  >> A broader range is generally better.

- What is the gender, age and racial sampling of the data?
  >> Greater balance on these dimensions is generally better.

Such questions are good and important, although many papers are published without such properties of the dataset being detailed.

WHAT? Typically we focus on the “what” elements of the dataset.

HOW? Nevertheless, the generalizability issue should also raise questions about “how” the images are used to estimate accuracy, as well as “what” is represented in the images.

HOW?

- How is the accuracy estimated?
  >> Mean? weighted mean?

- How is the experimental protocol defined?
  >>Leave-one-out? Half-Half? 10-fold cross-validation?

- How are the images divided into train and test portions?
  >> Randomly? Every *N*-th image?
  >> According to time of acquisition?

- How is the data sampled from the underlying original data collection?
  >> Is any data that was originally collected not used?
  >> If so, is this documented?

- How is the person-specific nature of the data captured?
  >>Are train and test splits person-disjoint?

- How is the variance in the estimated accuracy estimated?
  >> High? Low?

A real example about ‘how’…

A

SAMPLE TITLE OF SAMPLE PAPER WITH MORE THAN ONE AUTHOR

The images were divided into ten subsets. For each validation fold, nine subsets were used for training purposes and the remaining subset was used for testing. This step was repeated ten times, and the accuracy of the ten folds were averaged to compute the final estimation of the accuracy. The classification rate was 96.3%.

B

Software Tools: A Key Component in the Successful Implementation of the ATML Standards

The subjects used for training were not used for testing. We have used a 10-fold cross validation procedure. The estimated accuracy was 70.0%

C

Bridging the Gap between Social Software and Business Process Management: A Research Agenda

We divide 10 facial expression sequences of every person into training and testing sets. Firstly, we use one expression image for testing, others for training. Then 14 images are used for training and 7 images left for testing. At last 7 images are used for training and 14 images for testing. The recognition rate can reach about 95.0%.

Typical Problems

There are two typical problems in “how” the images were used in the experiments:

1. no standard protocol, and
2. ill-defined protocols.

They undermine the research on biometrics because they lead to confusing differences in strength of protocol with differences in estimated accuracy of algorithms.

We propose the ‘EPD Methodology’…

E - Experiments

Wherever a subject-disjoint train-and-test split would be possible, it should be used.

P - Protocol

The protocol should ideally be to report the mean and standard deviation of some number of randomized 10-fold cross validation trials.

Reviewers should accept accuracy reported on a single hold out trial only if there is a clear justification made.

D - Data

Any downsampling from the collected dataset should be described and justified. Wherever possible, results should be presented with and without the downsampling, so that reviewers can judge its effect.

Toy Example

[SIMULATED DATA]

2 Classes

2 Gaussian Distributions

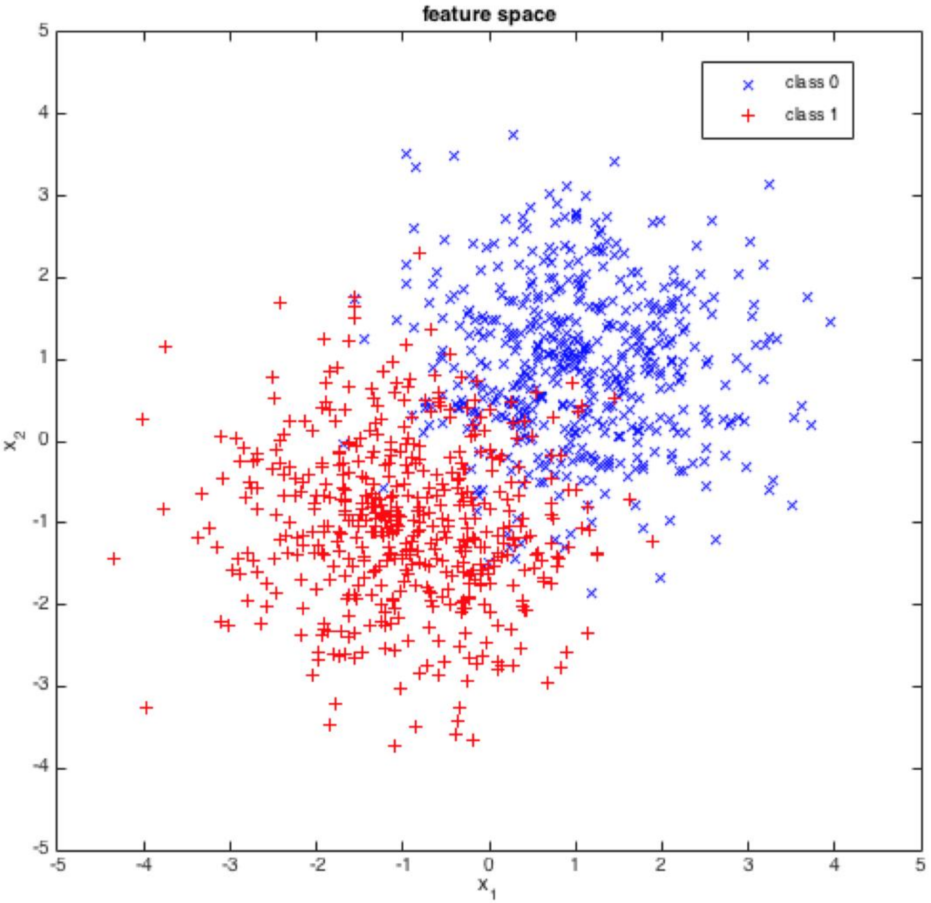
$\mu_1 = (1,1)$

$\mu_2 = (-1,-1)$

$\sigma_1 = \sigma_1 = 1$

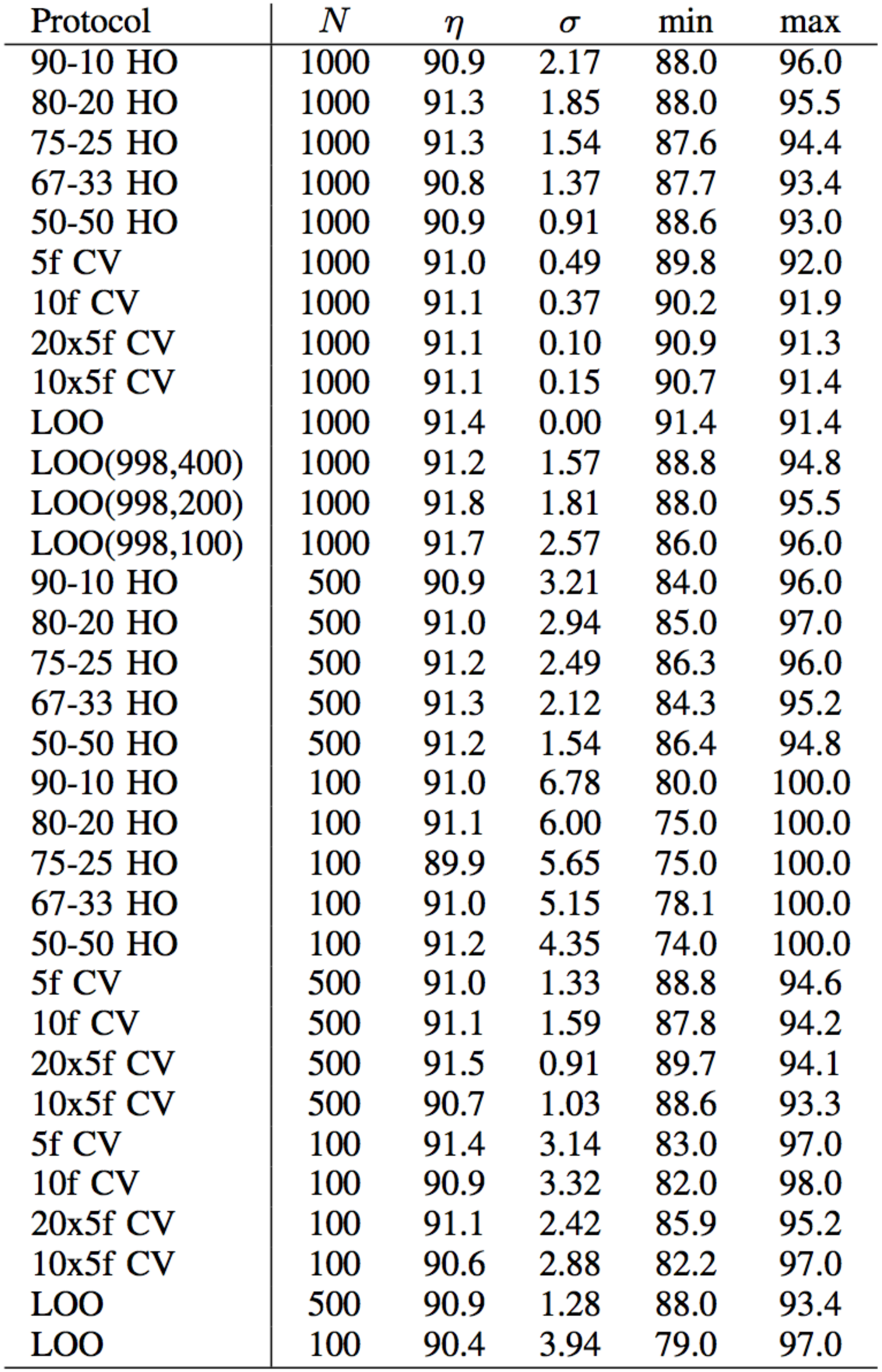
500 samples /class

$N_0 = 1000$ (available data)



Accuracy Estimation using KNN (K=3)

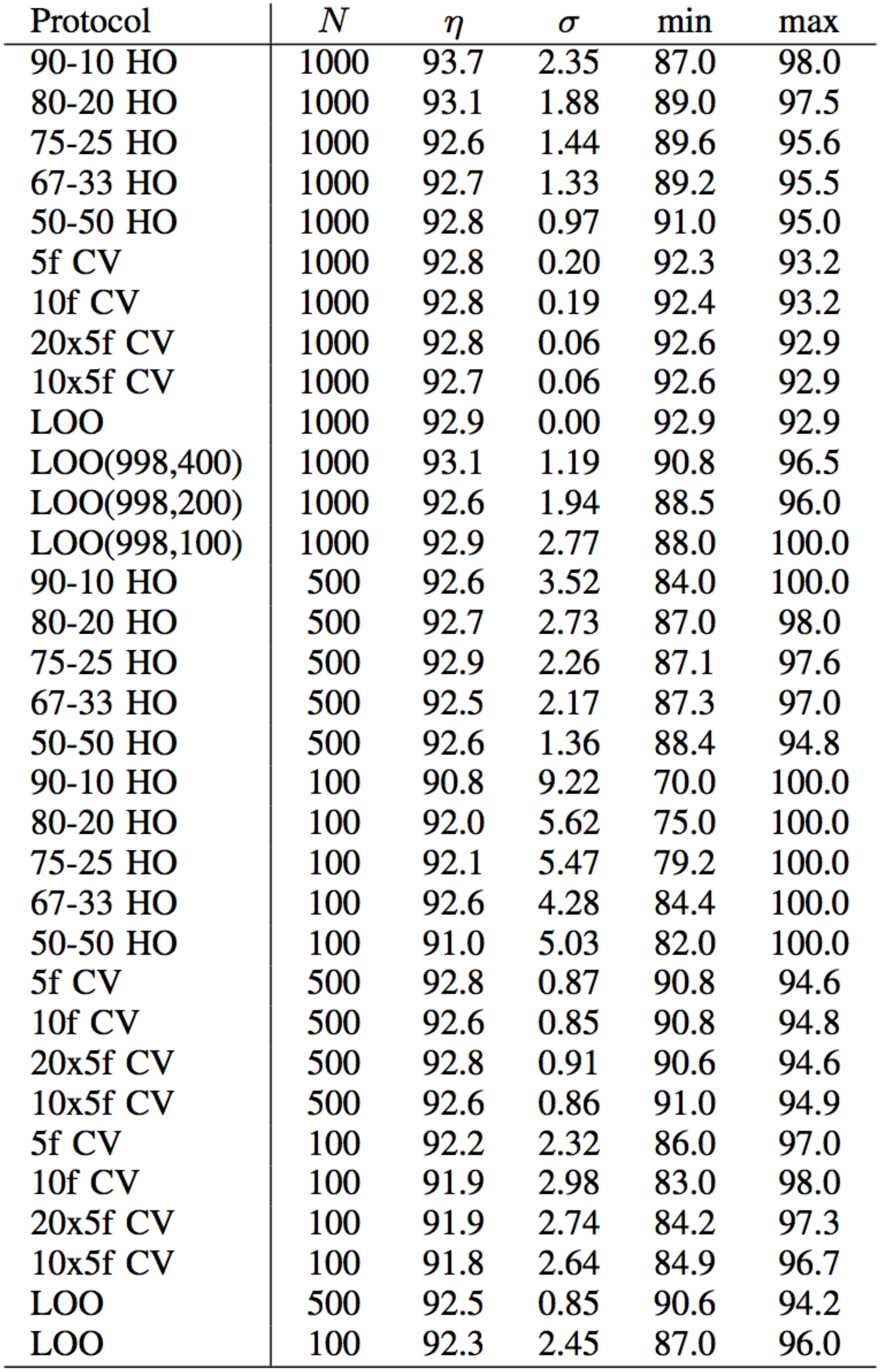
AFTER 50 REPETITIONS



Protocol	N	η	σ	min	max
90-10 HO	1000	90.9	2.17	88.0	96.0
80-20 HO	1000	91.3	1.85	88.0	95.5
75-25 HO	1000	91.3	1.54	87.6	94.4
67-33 HO	1000	90.8	1.37	87.7	93.4
50-50 HO	1000	90.9	0.91	88.6	93.0
5f CV	1000	91.0	0.49	89.8	92.0
10f CV	1000	91.1	0.37	90.2	91.9
20x5f CV	1000	91.1	0.10	90.9	91.3
10x5f CV	1000	91.1	0.15	90.7	91.4
LOO	1000	91.4	0.00	91.4	91.4
LOO(998,400)	1000	91.2	1.57	88.8	94.8
LOO(998,200)	1000	91.8	1.81	88.0	95.5
LOO(998,100)	1000	91.7	2.57	86.0	96.0
90-10 HO	500	90.9	3.21	84.0	96.0
80-20 HO	500	91.0	2.94	85.0	97.0
75-25 HO	500	91.2	2.49	86.3	96.0
67-33 HO	500	91.3	2.12	84.3	95.2
50-50 HO	500	91.2	1.54	86.4	94.8
90-10 HO	100	91.0	6.78	80.0	100.0
80-20 HO	100	91.1	6.00	75.0	100.0
75-25 HO	100	89.9	5.65	75.0	100.0
67-33 HO	100	91.0	5.15	78.1	100.0
50-50 HO	100	91.2	4.35	74.0	100.0
5f CV	500	91.0	1.33	88.8	94.6
10f CV	500	91.1	1.59	87.8	94.2
20x5f CV	500	91.5	0.91	89.7	94.1
10x5f CV	500	90.7	1.03	88.6	93.3
5f CV	100	91.4	3.14	83.0	97.0
10f CV	100	90.9	3.32	82.0	98.0
20x5f CV	100	91.1	2.42	85.9	95.2
10x5f CV	100	90.6	2.88	82.2	97.0
LOO	500	90.9	1.28	88.0	93.4
LOO	100	90.4	3.94	79.0	97.0

Accuracy Estimation using Linear-SVM

AFTER 50 REPETITIONS

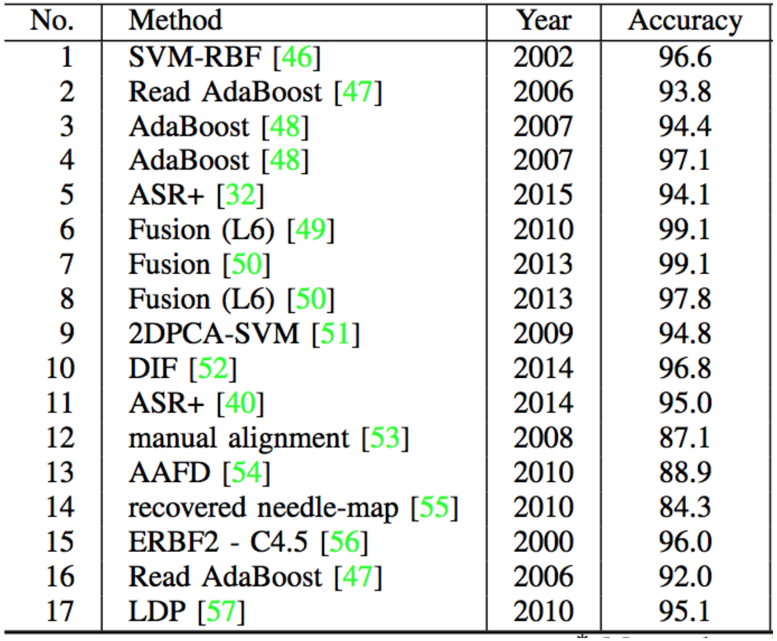


Protocol	N	η	σ	min	max
90-10 HO	1000	93.7	2.35	87.0	98.0
80-20 HO	1000	93.1	1.88	89.0	97.5
75-25 HO	1000	92.6	1.44	89.6	95.6
67-33 HO	1000	92.7	1.33	89.2	95.5
50-50 HO	1000	92.8	0.97	91.0	95.0
5f CV	1000	92.8	0.20	92.3	93.2
10f CV	1000	92.8	0.19	92.4	93.2
20x5f CV	1000	92.8	0.06	92.6	92.9
10x5f CV	1000	92.7	0.06	92.6	92.9
LOO	1000	92.9	0.00	92.9	92.9
LOO(998,400)	1000	93.1	1.19	90.8	96.5
LOO(998,200)	1000	92.6	1.94	88.5	96.0
LOO(998,100)	1000	92.9	2.77	88.0	100.0
90-10 HO	500	92.6	3.52	84.0	100.0
80-20 HO	500	92.7	2.73	87.0	98.0
75-25 HO	500	92.9	2.26	87.1	97.6
67-33 HO	500	92.5	2.17	87.3	97.0
50-50 HO	500	92.6	1.36	88.4	94.8
90-10 HO	100	90.8	9.22	70.0	100.0
80-20 HO	100	92.0	5.62	75.0	100.0
75-25 HO	100	92.1	5.47	79.2	100.0
67-33 HO	100	92.6	4.28	84.4	100.0
50-50 HO	100	91.0	5.03	82.0	100.0
5f CV	500	92.8	0.87	90.8	94.6
10f CV	500	92.6	0.85	90.8	94.8
20x5f CV	500	92.8	0.91	90.6	94.6
10x5f CV	500	92.6	0.86	91.0	94.9
5f CV	100	92.2	2.32	86.0	97.0
10f CV	100	91.9	2.98	83.0	98.0
20x5f CV	100	91.9	2.74	84.2	97.3
10x5f CV	100	91.8	2.64	84.9	96.7
LOO	500	92.5	0.85	90.6	94.2
LOO	100	92.3	2.45	87.0	96.0

Literature Review

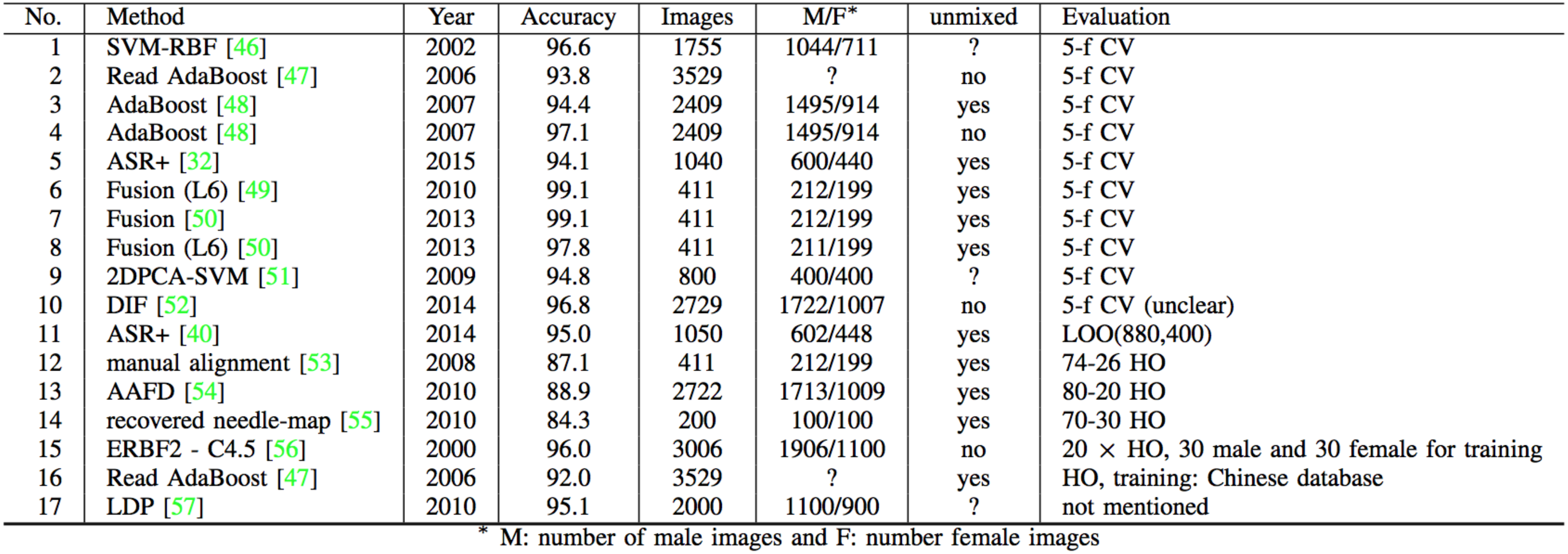
FERET
Gender

Gender Classification using FERET Dataset



No.	Method	Year	Accuracy
1	SVM-RBF [46]	2002	96.6
2	Read AdaBoost [47]	2006	93.8
3	AdaBoost [48]	2007	94.4
4	AdaBoost [48]	2007	97.1
5	ASR+ [32]	2015	94.1
6	Fusion (L6) [49]	2010	99.1
7	Fusion [50]	2013	99.1
8	Fusion (L6) [50]	2013	97.8
9	2DPCA-SVM [51]	2009	94.8
10	DIF [52]	2014	96.8
11	ASR+ [40]	2014	95.0
12	manual alignment [53]	2008	87.1
13	AAFD [54]	2010	88.9
14	recovered needle-map [55]	2010	84.3
15	ERBF2 - C4.5 [56]	2000	96.0
16	Read AdaBoost [47]	2006	92.0
17	LDP [57]	2010	95.1
]

Gender Classification using FERET Dataset



No.	Method	Year	Accuracy	Images	M/F*	unmixed	Evaluation
1	SVM-RBF [46]	2002	96.6	1755	1044/711	?	5-f CV
2	Read AdaBoost [47]	2006	93.8	3529	?	no	5-f CV
3	AdaBoost [48]	2007	94.4	2409	1495/914	yes	5-f CV
4	AdaBoost [48]	2007	97.1	2409	1495/914	no	5-f CV
5	ASR+ [32]	2015	94.1	1040	600/440	yes	5-f CV
6	Fusion (L6) [49]	2010	99.1	411	212/199	yes	5-f CV
7	Fusion [50]	2013	99.1	411	212/199	yes	5-f CV
8	Fusion (L6) [50]	2013	97.8	411	211/199	yes	5-f CV
9	2DPCA-SVM [51]	2009	94.8	800	400/400	?	5-f CV
10	DIF [52]	2014	96.8	2729	1722/1007	no	5-f CV (unclear)
11	ASR+ [40]	2014	95.0	1050	602/448	yes	LOO(880,400)
12	manual alignment [53]	2008	87.1	411	212/199	yes	74-26 HO
13	AAFD [54]	2010	88.9	2722	1713/1009	yes	80-20 HO
14	recovered needle-map [55]	2010	84.3	200	100/100	yes	70-30 HO
15	ERBF2 - C4.5 [56]	2000	96.0	3006	1906/1100	no	20 × HO, 30 male and 30 female for training
16	Read AdaBoost [47]	2006	92.0	3529	?	yes	HO, training: Chinese database
17	LDP [57]	2010	95.1	2000	1100/900	?	not mentioned

* M: number of male images and F: number female images

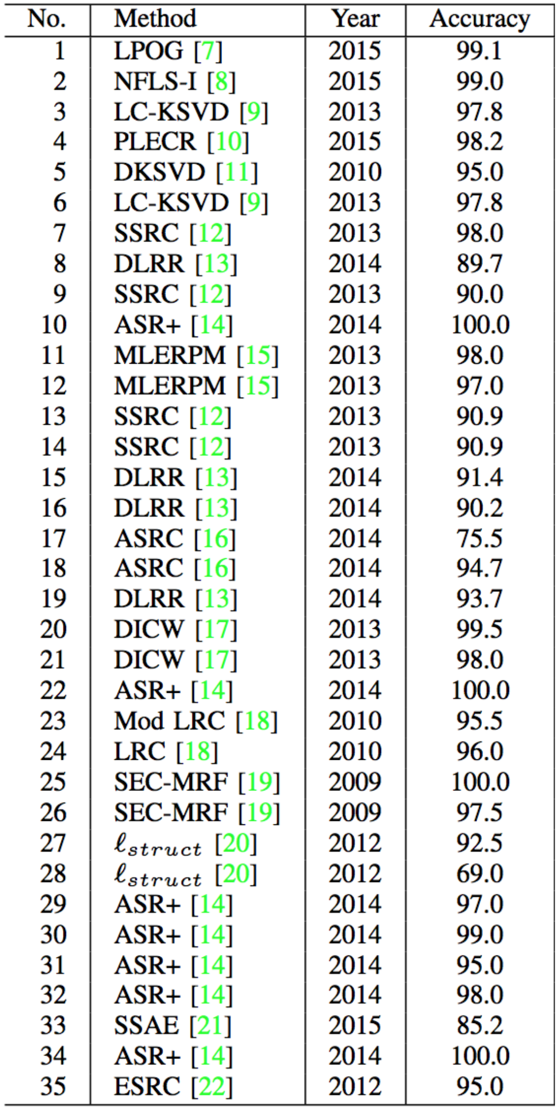
14 papers with 17 gender classification results using the FERET dataset.

How many pairs of papers have directly comparable results?

One.

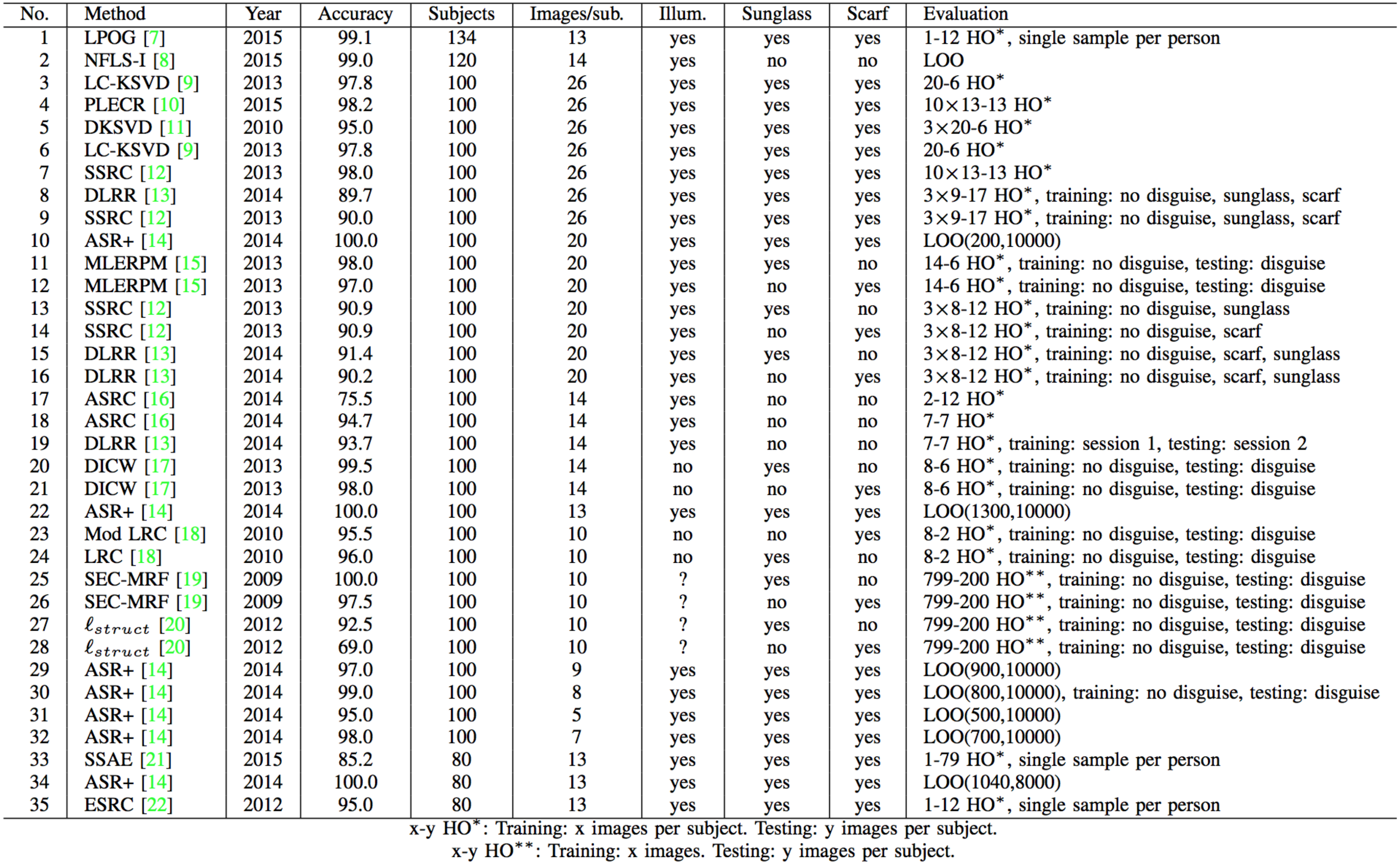
AR
Face Recognition

Face recognition using AR Dataset



No.	Method	Year	Accuracy
1	LPOG [7]	2015	99.1
2	NFLS-I [8]	2015	99.0
3	LC-KSVD [9]	2013	97.8
4	PLECR [10]	2015	98.2
5	DKSVD [11]	2010	95.0
6	LC-KSVD [9]	2013	97.8
7	SSRC [12]	2013	98.0
8	DLRR [13]	2014	89.7
9	SSRC [12]	2013	90.0
10	ASR+ [14]	2014	100.0
11	MLERPM [15]	2013	98.0
12	MLERPM [15]	2013	97.0
13	SSRC [12]	2013	90.9
14	SSRC [12]	2013	90.9
15	DLRR [13]	2014	91.4
16	DLRR [13]	2014	90.2
17	ASRC [16]	2014	75.5
18	ASRC [16]	2014	94.7
19	DLRR [13]	2014	93.7
20	DICW [17]	2013	99.5
21	DICW [17]	2013	98.0
22	ASR+ [14]	2014	100.0
23	Mod LRC [18]	2010	95.5
24	LRC [18]	2010	96.0
25	SEC-MRF [19]	2009	100.0
26	SEC-MRF [19]	2009	97.5
27	ℓ_{struct} [20]	2012	92.5
28	ℓ_{struct} [20]	2012	69.0
29	ASR+ [14]	2014	97.0
30	ASR+ [14]	2014	99.0
31	ASR+ [14]	2014	95.0
32	ASR+ [14]	2014	98.0
33	SSAE [21]	2015	85.2
34	ASR+ [14]	2014	100.0
35	ESRC [22]	2012	95.0

Face recognition using AR Dataset



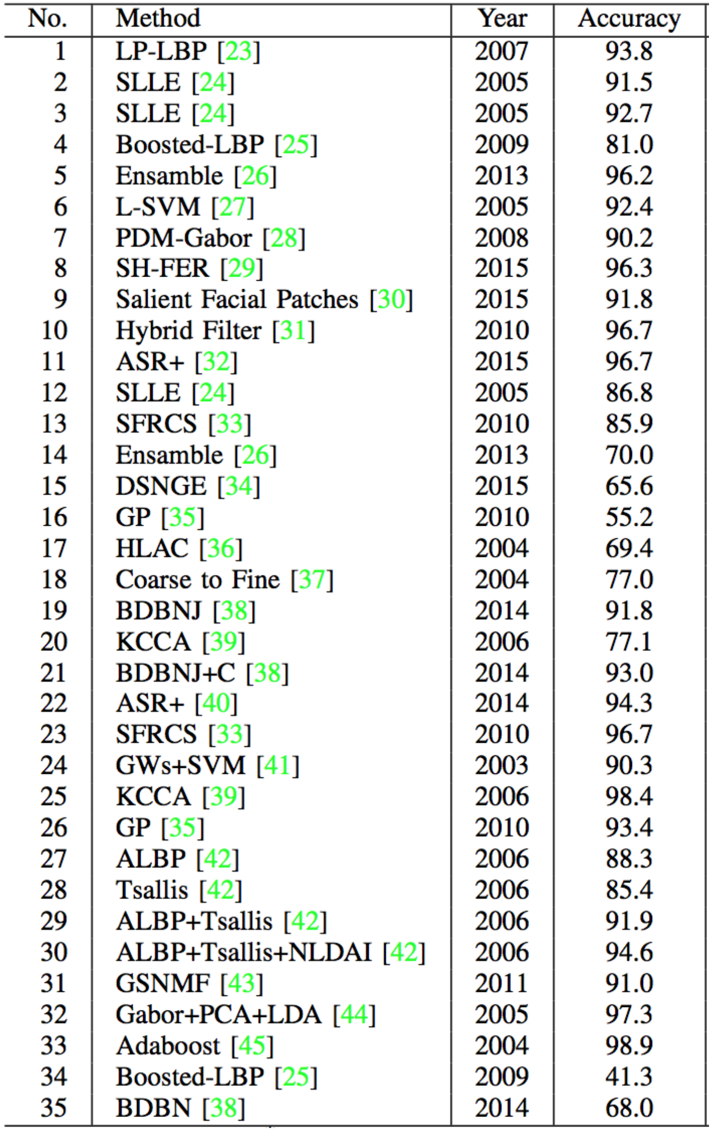
No.	Method	Year	Accuracy	Subjects	Images/sub.	Illum.	Sunglass	Scarf	Evaluation
1	LPOG [7]	2015	99.1	134	13	yes	yes	yes	1-12 HO*, single sample per person
2	NFLS-I [8]	2015	99.0	120	14	yes	no	no	LOO
3	LC-KSVD [9]	2013	97.8	100	26	yes	yes	yes	20-6 HO*
4	PLECR [10]	2015	98.2	100	26	yes	yes	yes	10×13-13 HO*
5	DKSVD [11]	2010	95.0	100	26	yes	yes	yes	3×20-6 HO*
6	LC-KSVD [9]	2013	97.8	100	26	yes	yes	yes	20-6 HO*
7	SSRC [12]	2013	98.0	100	26	yes	yes	yes	10×13-13 HO*
8	DLRR [13]	2014	89.7	100	26	yes	yes	yes	3×9-17 HO*, training: no disguise, sunglass, scarf
9	SSRC [12]	2013	90.0	100	26	yes	yes	yes	3×9-17 HO*, training: no disguise, sunglass, scarf
10	ASR+ [14]	2014	100.0	100	20	yes	yes	yes	LOO(200,10000)
11	MLERPM [15]	2013	98.0	100	20	yes	yes	no	14-6 HO*, training: no disguise, testing: disguise
12	MLERPM [15]	2013	97.0	100	20	yes	no	yes	14-6 HO*, training: no disguise, testing: disguise
13	SSRC [12]	2013	90.9	100	20	yes	yes	no	3×8-12 HO*, training: no disguise, sunglass
14	SSRC [12]	2013	90.9	100	20	yes	no	yes	3×8-12 HO*, training: no disguise, scarf
15	DLRR [13]	2014	91.4	100	20	yes	yes	no	3×8-12 HO*, training: no disguise, scarf, sunglass
16	DLRR [13]	2014	90.2	100	20	yes	no	yes	3×8-12 HO*, training: no disguise, scarf, sunglass
17	ASRC [16]	2014	75.5	100	14	yes	no	no	2-12 HO*
18	ASRC [16]	2014	94.7	100	14	yes	no	no	7-7 HO*
19	DLRR [13]	2014	93.7	100	14	yes	no	no	7-7 HO*, training: session 1, testing: session 2
20	DICW [17]	2013	99.5	100	14	no	yes	no	8-6 HO*, training: no disguise, testing: disguise
21	DICW [17]	2013	98.0	100	14	no	no	yes	8-6 HO*, training: no disguise, testing: disguise
22	ASR+ [14]	2014	100.0	100	13	yes	yes	yes	LOO(1300,10000)
23	Mod LRC [18]	2010	95.5	100	10	no	no	yes	8-2 HO*, training: no disguise, testing: disguise
24	LRC [18]	2010	96.0	100	10	no	yes	no	8-2 HO*, training: no disguise, testing: disguise
25	SEC-MRF [19]	2009	100.0	100	10	?	yes	no	799-200 HO**, training: no disguise, testing: disguise
26	SEC-MRF [19]	2009	97.5	100	10	?	no	yes	799-200 HO**, training: no disguise, testing: disguise
27	ℓ_{struct} [20]	2012	92.5	100	10	?	yes	no	799-200 HO**, training: no disguise, testing: disguise
28	ℓ_{struct} [20]	2012	69.0	100	10	?	no	yes	799-200 HO**, training: no disguise, testing: disguise
29	ASR+ [14]	2014	97.0	100	9	yes	yes	yes	LOO(900,10000)
30	ASR+ [14]	2014	99.0	100	8	yes	yes	yes	LOO(800,10000), training: no disguise, testing: disguise
31	ASR+ [14]	2014	95.0	100	5	yes	yes	yes	LOO(500,10000)
32	ASR+ [14]	2014	98.0	100	7	yes	yes	yes	LOO(700,10000)
33	SSAE [21]	2015	85.2	80	13	yes	yes	yes	1-79 HO*, single sample per person
34	ASR+ [14]	2014	100.0	80	13	yes	yes	yes	LOO(1040,8000)
35	ESRC [22]	2012	95.0	80	13	yes	yes	yes	1-12 HO*, single sample per person

x-y HO*: Training: x images per subject. Testing: y images per subject.

x-y HO**: Training: x images. Testing: y images per subject.

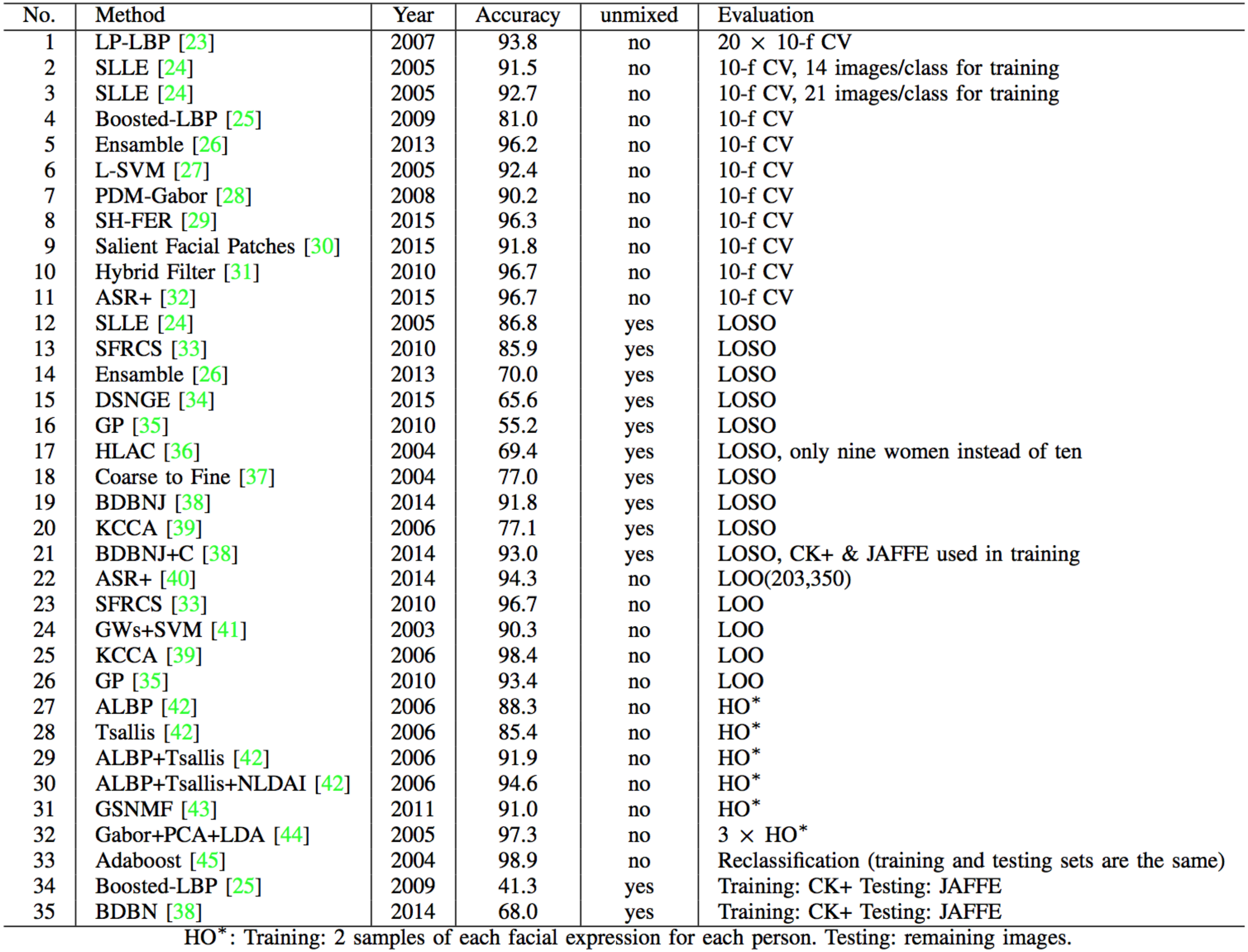
JAFFE
Expressions

Expression recognition using JAFFE Dataset



No.	Method	Year	Accuracy
1	LP-LBP [23]	2007	93.8
2	SLLE [24]	2005	91.5
3	SLLE [24]	2005	92.7
4	Boosted-LBP [25]	2009	81.0
5	Ensamble [26]	2013	96.2
6	L-SVM [27]	2005	92.4
7	PDM-Gabor [28]	2008	90.2
8	SH-FER [29]	2015	96.3
9	Salient Facial Patches [30]	2015	91.8
10	Hybrid Filter [31]	2010	96.7
11	ASR+ [32]	2015	96.7
12	SLLE [24]	2005	86.8
13	SFRCS [33]	2010	85.9
14	Ensamble [26]	2013	70.0
15	DSNGE [34]	2015	65.6
16	GP [35]	2010	55.2
17	HLAC [36]	2004	69.4
18	Coarse to Fine [37]	2004	77.0
19	BDBNJ [38]	2014	91.8
20	KCCA [39]	2006	77.1
21	BDBNJ+C [38]	2014	93.0
22	ASR+ [40]	2014	94.3
23	SFRCS [33]	2010	96.7
24	GWs+SVM [41]	2003	90.3
25	KCCA [39]	2006	98.4
26	GP [35]	2010	93.4
27	ALBP [42]	2006	88.3
28	Tsallis [42]	2006	85.4
29	ALBP+Tsallis [42]	2006	91.9
30	ALBP+Tsallis+NLDAI [42]	2006	94.6
31	GSNMF [43]	2011	91.0
32	Gabor+PCA+LDA [44]	2005	97.3
33	Adaboost [45]	2004	98.9
34	Boosted-LBP [25]	2009	41.3
35	BDBN [38]	2014	68.0

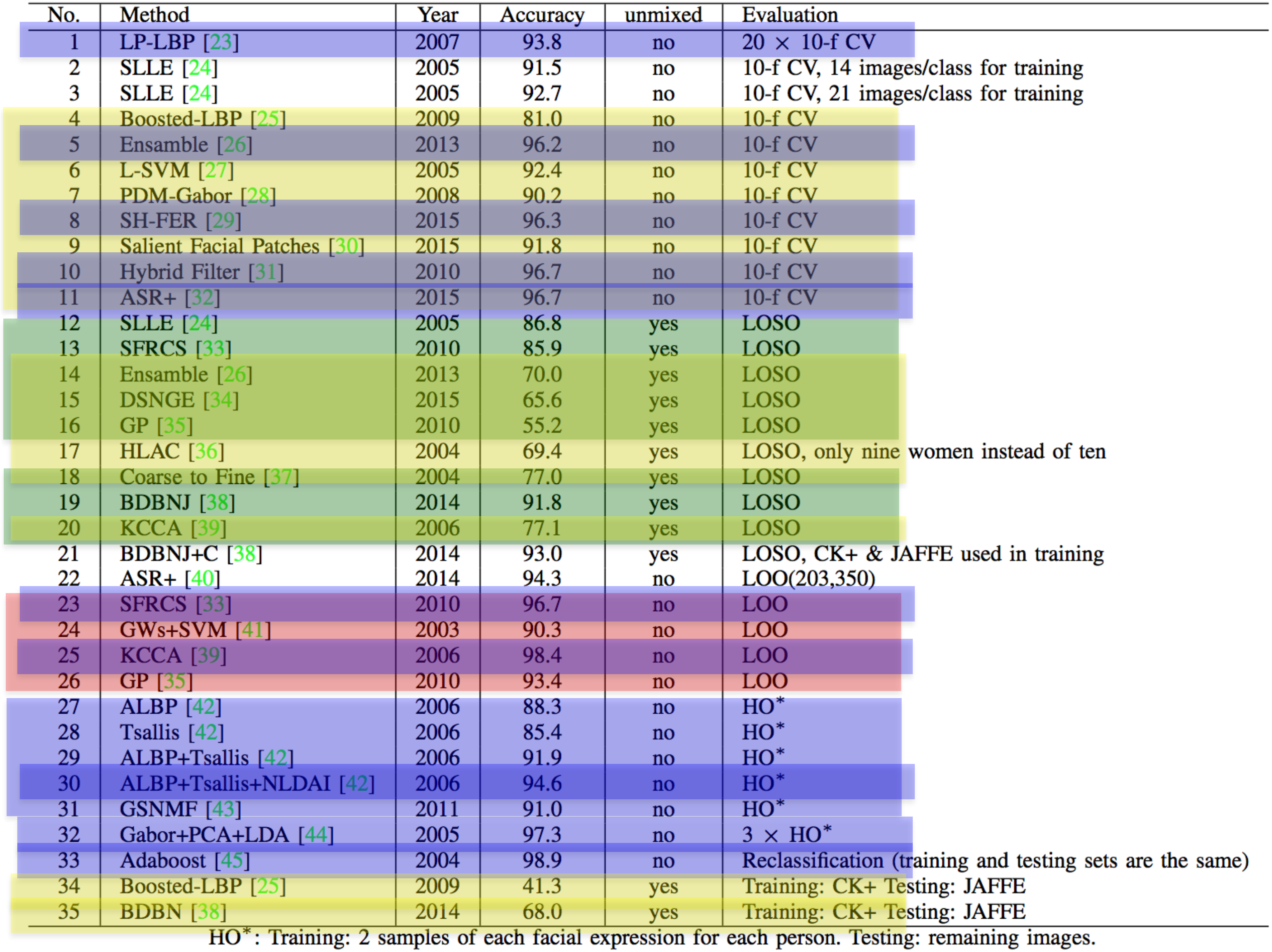
Expression recognition using JAFFE Dataset



No.	Method	Year	Accuracy	unmixed	Evaluation
1	LP-LBP [23]	2007	93.8	no	20 × 10-f CV
2	SLLE [24]	2005	91.5	no	10-f CV, 14 images/class for training
3	SLLE [24]	2005	92.7	no	10-f CV, 21 images/class for training
4	Boosted-LBP [25]	2009	81.0	no	10-f CV
5	Ensamble [26]	2013	96.2	no	10-f CV
6	L-SVM [27]	2005	92.4	no	10-f CV
7	PDM-Gabor [28]	2008	90.2	no	10-f CV
8	SH-FER [29]	2015	96.3	no	10-f CV
9	Salient Facial Patches [30]	2015	91.8	no	10-f CV
10	Hybrid Filter [31]	2010	96.7	no	10-f CV
11	ASR+ [32]	2015	96.7	no	10-f CV
12	SLLE [24]	2005	86.8	yes	LOSO
13	SFRCS [33]	2010	85.9	yes	LOSO
14	Ensamble [26]	2013	70.0	yes	LOSO
15	DSNGE [34]	2015	65.6	yes	LOSO
16	GP [35]	2010	55.2	yes	LOSO
17	HLAC [36]	2004	69.4	yes	LOSO, only nine women instead of ten
18	Coarse to Fine [37]	2004	77.0	yes	LOSO
19	BDBNJ [38]	2014	91.8	yes	LOSO
20	KCCA [39]	2006	77.1	yes	LOSO
21	BDBNJ+C [38]	2014	93.0	yes	LOSO, CK+ & JAFFE used in training
22	ASR+ [40]	2014	94.3	no	LOO(203,350)
23	SFRCS [33]	2010	96.7	no	LOO
24	GWs+SVM [41]	2003	90.3	no	LOO
25	KCCA [39]	2006	98.4	no	LOO
26	GP [35]	2010	93.4	no	LOO
27	ALBP [42]	2006	88.3	no	HO*
28	Tsallis [42]	2006	85.4	no	HO*
29	ALBP+Tsallis [42]	2006	91.9	no	HO*
30	ALBP+Tsallis+NLDAI [42]	2006	94.6	no	HO*
31	GSNMF [43]	2011	91.0	no	HO*
32	Gabor+PCA+LDA [44]	2005	97.3	no	3 × HO*
33	Adaboost [45]	2004	98.9	no	Reclassification (training and testing sets are the same)
34	Boosted-LBP [25]	2009	41.3	yes	Training: CK+ Testing: JAFFE
35	BDBN [38]	2014	68.0	yes	Training: CK+ Testing: JAFFE

HO*: Training: 2 samples of each facial expression for each person. Testing: remaining images.

Expression recognition using JAFFE Dataset



No.	Method	Year	Accuracy	unmixed	Evaluation
1	LP-LBP [23]	2007	93.8	no	20 × 10-f CV
2	SLLE [24]	2005	91.5	no	10-f CV, 14 images/class for training
3	SLLE [24]	2005	92.7	no	10-f CV, 21 images/class for training
4	Boosted-LBP [25]	2009	81.0	no	10-f CV
5	Ensamble [26]	2013	96.2	no	10-f CV
6	L-SVM [27]	2005	92.4	no	10-f CV
7	PDM-Gabor [28]	2008	90.2	no	10-f CV
8	SH-FER [29]	2015	96.3	no	10-f CV
9	Salient Facial Patches [30]	2015	91.8	no	10-f CV
10	Hybrid Filter [31]	2010	96.7	no	10-f CV
11	ASR+ [32]	2015	96.7	no	10-f CV
12	SLLE [24]	2005	86.8	yes	LOSO
13	SFRCS [33]	2010	85.9	yes	LOSO
14	Ensamble [26]	2013	70.0	yes	LOSO
15	DSNGE [34]	2015	65.6	yes	LOSO
16	GP [35]	2010	55.2	yes	LOSO
17	HLAC [36]	2004	69.4	yes	LOSO, only nine women instead of ten
18	Coarse to Fine [37]	2004	77.0	yes	LOSO
19	BDBNJ [38]	2014	91.8	yes	LOSO
20	KCCA [39]	2006	77.1	yes	LOSO
21	BDBNJ+C [38]	2014	93.0	yes	LOSO, CK+ & JAFFE used in training
22	ASR+ [40]	2014	94.3	no	LOO(203,350)
23	SFRCS [33]	2010	96.7	no	LOO
24	GWs+SVM [41]	2003	90.3	no	LOO
25	KCCA [39]	2006	98.4	no	LOO
26	GP [35]	2010	93.4	no	LOO
27	ALBP [42]	2006	88.3	no	HO*
28	Tsallis [42]	2006	85.4	no	HO*
29	ALBP+Tsallis [42]	2006	91.9	no	HO*
30	ALBP+Tsallis+NLDAI [42]	2006	94.6	no	HO*
31	GSNMF [43]	2011	91.0	no	HO*
32	Gabor+PCA+LDA [44]	2005	97.3	no	3 × HO*
33	Adaboost [45]	2004	98.9	no	Reclassification (training and testing sets are the same)
34	Boosted-LBP [25]	2009	41.3	yes	Training: CK+ Testing: JAFFE
35	BDBN [38]	2014	68.0	yes	Training: CK+ Testing: JAFFE

HO*: Training: 2 samples of each facial expression for each person. Testing: remaining images.

Conclusions

Conclusions

- Using the same algorithm and dataset, the estimated accuracy can be totally different depending on
  - the selection of training and testing data,
  - the number of samples of the used dataset and
  - the number of single accuracies averaged to estimate accuracy

- Based on the published literature, it is rare to find two papers published on the same problem that use the same experimental protocol in all important elements.

Conclusions

- For problems where a subject-disjoint train-and-test split is essential in order to obtain a useful accuracy estimate, papers are often published using a non-disjoint split.

- For problems of this type, a leave-one-subject-out protocol would seem to be the default recommendation for useful experimental results.

Conclusions

- The quality of experimental results in the biometrics literature could be improved if authors, reviewers and editors followed EPD methodology:
  - **E**xperiments: if subject-disjoint is appropriate and is not used, then recommend reject.
  - **P**rotocol: If single-trial HO is used, recommend reject. If dataset is large enough, single-trial CV might be sufficient; average of N trials is better.
  - **D**ataset: Selection from dataset must be justified in context. Present results with and without.

On Accuracy Estimation in Face Biometric Problems

Domingo Mery, Yuning Zhao and Kevin Bowyer

UNIVERSITY OF NOTRE DAME
College of Engineering
UNIVERSITY OF NOTRE DAME

FACULTAD DE INGENIERÍA
PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE